https://doi.org/10.1038/s44260-024-00002-2

Adaptive link dynamics drive online hate networks and their mainstream influence

Minzhang Zheng[1], Richard F. Sear[1], Lucia Illari[1], Nicholas J. Restrepo[2] & Neil F. Johnson[1]

Online hate is dynamic, adaptive— and may soon surge with new AI/GPT tools. Establishing how hate operates at scale is key to overcoming it. We provide insights that challenge existing policies. Rather than large social media platforms being the key drivers, waves of adaptive links across smaller platforms connect the hate user base over time, fortifying hate networks, bypassing mitigations, and extending their direct influence into the massive neighboring mainstream. Data indicates that hundreds of thousands of people globally, including children, have been exposed. We present governing equations derived from first principles and a tipping-point condition predicting future surges in content transmission. Using the U.S. Capitol attack and a 2023 mass shooting as case studies, our findings offer actionable insights and quantitative predictions down to the hourly scale. The efficacy of proposed mitigations can now be predicted using these equations.

There is no accepted scientific understanding of how online hate manages to thrive at scale. Yet its societal consequences are widespread and occur daily globally, e.g., personal traumas[1–3]; gender, race, and religion-based abuse; child sex abuse[4,5]; and violent mass attacks[6]. Making matters more complex from a scientific viewpoint, online hate is dynamic, adaptive[7–11] and now looks set to surge[12–14] armed with new AI/GPT tools[15,16]. Overcoming it will require establishing the science of how it operates at scale[17].

In addition to the obvious consequences for the direct victims of hate attacks and abuse, there is a huge secondary impact. Nearly 50% of all Americans now compromise aspects of their and their children's daily lives in order to lower the risk of experiencing some hate-driven mass shooting, e.g., 6 May 2023 Allen, Texas shooting which appears to be one of an increasing number inspired by social media hate content[18,19]. Separately, 2024 will see more than 60 elections across 54 countries including the U.S. and India, where the scope for online hate to cause voter intimidation is huge[20,21]. Such mass-scale threats, now supercharged by AI/GPT weaponry, are accelerating efforts to win the war against online hate and other harms[22–37]. We refer to refs. 38–40 for unifying perspectives on this huge and still growing body of research, while ref. 41 provides daily updates on new studies.

The current war against online hate of all forms is being led on the regulatory side by the EU's "Digital Services Act" (DSA) and "A.I. Act"[42,43]. Social media platforms (referred to hereafter as "platforms") on the list of "Very Large Online Platforms" such as Facebook and Twitter, must carry out a risk assessment which includes an analysis of how harmful content might be disseminated through their service[44]. At face value, this appears to make perfect sense since the largest platforms (e.g., Facebook, Twitter) have the largest share of users. Hateful content is thought to occupy the fringes of the Internet[45–50]. However, winning any war requires an accurate picture of the battlefield.

Here we adopt an engineering approach and show how it leads to an understanding of the machinery and link dynamics that manage to sustain online hate at scale. We start by mapping out the dynamical structure of the online hate network across platforms. Instead of the large platforms being the key drivers, we find that waves of adaptive links connect the hate user base over time across a large ecosystem of smaller platforms, allowing hate networks to steadily strengthen, bypass mitigations, and increase their direct influence on the massive neighboring mainstream. The data suggests hundreds of thousands of individuals globally have recently been exposed, including children. We then establish governing dynamical equations derived from first principles. A tipping-point condition predicts more frequent future surges in content transmission.

Using the 2021 U.S. Capitol attack and a 2023 mass shooting as illustrations, we show our findings offer abiding insights and quantitative predictions down to the hourly scale. This dynamical focus yields general information about online hate networks, their connection to the mainstream population, the question of when information will/will not be transmitted at a global scale, and the effects of mitigation strategies.

To improve the flow of the paper for a general audience, we have put the full technical details of the data collection as well as the mathematical derivations and calculations, into the Supplementary Information (SI) which accompanies the paper online. Figures 1–3 show the key results.

In this paper, we frequently refer to "links," which can be understood as edges in the hate network. A link is formed between a source community

[1]Dynamic Online Networks Laboratory, George Washington University, Washington, D.C. 20052, USA. [2]ClustrX LLC, Washington, D.C., USA.
e-mail: neiljohnson@gwu.edu

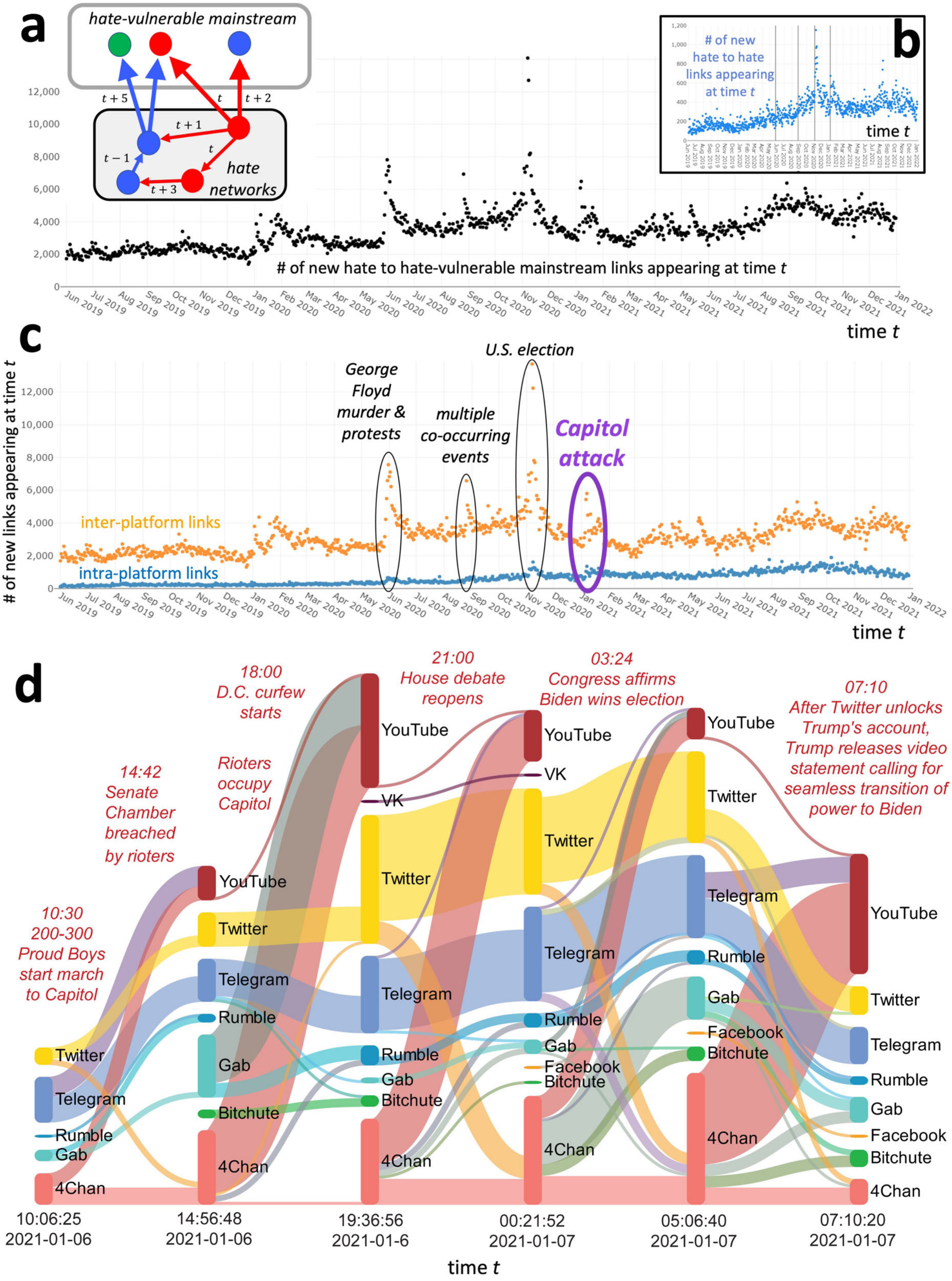

Fig. 1 | Hate's highly adaptive link dynamics. a–c Number of new links created on each day t from hate communities (nodes). Time-series show different aggregations across the 4 link types (schematic in panel a left inset): hate to hate inter-platform; hate to hate intra-platform; hate to hate-vulnerable mainstream inter-platform; hate to hate-vulnerable mainstream intra-platform. Suppl. Fig. 4 gives explicit examples of these links. **d** Sankey diagram shows intra-day flows of new links from hate nodes on a given platform (source) into hate nodes on a target platform. SI Sec. 3 explains Sankey diagram construction.

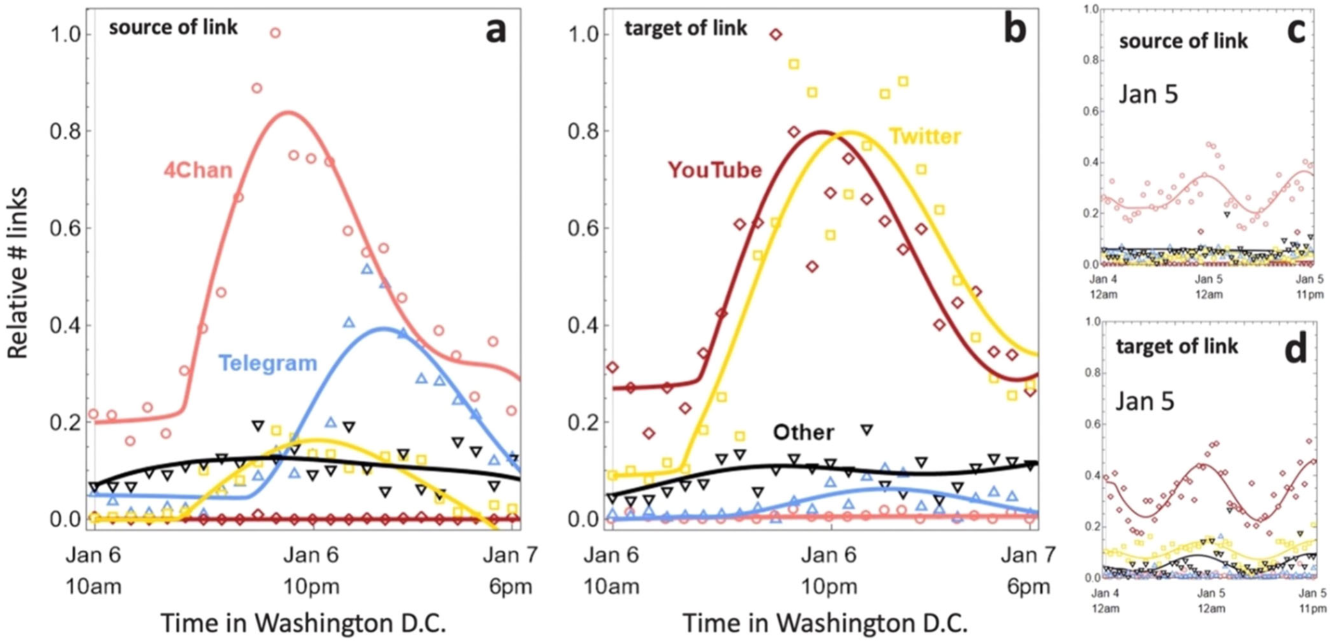

Fig. 2 | Empirical data (symbols) vs. mathematical solutions of the deterministic governing equations (curves, derivation SI Sec. 4.1). a Relative number of links created at time *t* from hate communities on a given platform (source). **b** Relative number of links created at time *t* to other communities on a given platform (target). Approximately 80% of targets are hate-vulnerable mainstream communities. Only the largest curves are shown, the rest are aggregated as 'Other' (black curves). **c**, **d** Same as plots **a** and **b** but applied to Jan 4-5 data using different time points.

and target community when the source community shares a URL leading directly to the target or to content posted by the target. Source communities are always hate communities and target communities can be either hate or "hate-vulnerable" communities. Please see "Methods" and SI Sec. 1 for full details of our data collection and community classifications.

Results

Key Features

We map the online hate ecosystem using communities as nodes and links as edges. The resulting network (schematic Fig. 1a left inset; results Figs. 1–3; data collection methodology in SI) is a directed dynamical network with link wiring that can change quickly over time within and across platforms, and strong direct linkage from the hate networks to the massive hate-vulnerable mainstream. It contains 1542 hate communities (nodes) that in 2.5 years have created 285,378 links to each other and 2,832,128 links into 385,719 hate-vulnerable mainstream communities. The membership size for 907 of these hate communities is publicly available and averages to approximately 28,000 per community, so we can estimate there are roughly 43 M individuals involved in the hate core.

While the nodes (communities) are fairly constant over time, the link number increases massively every day and hence steadily strengthens the hate networks and their potential mainstream influence. Each new link (e.g., Figs. S1 and S5) means members of the source hate community (node) can immediately engage with the target community (node), pass hate content to it, and influence it.

Three key features emerge (illustrated in Figs. 1–3) that hate mitigations must account for in order to be effective. These insights can be directly incorporated into new legislation and/or content moderation policy:

(1) They must focus on activity (links) between platforms, particularly including the many smaller platforms as shown explicitly in Fig. 1d—and not just activity within the largest platforms. To keep their users safe, it is ineffective for individual platforms to only focus on themselves.

(2) They must be nimble enough to outpace rapid link creation dynamics (e.g. those shown in Figs. 1–2) down to the scale of minutes within a day—in particular, the huge waves of links which appear suddenly around notable events (Fig. 1), and which could further enflame hate, anger, and distrust during these events or their aftermath, possibly inciting new violent acts.

(3) They must avoid the existing "brute force" strategy of chasing down a queue of reported links. This is akin to a game of whack-a-mole, i.e., existing links can get buried below newer content and hence become less relevant while fresh, unreported links become the new focus. Old links can also get removed on purpose by the community member(s), or the piece of content they are in gets removed. This link loss also means that the most active pathways that hate content spreads though are changing all the time, hence mitigations to prevent system-wide spreading need to account for this.

Governing equations and analysis

Current approaches to mitigation and legislation do not account for the key empirical features (1)–(3) that emerged. However, by deriving a set of governing dynamical equations that explicitly capture (1)–(3), we can systematically incorporate them into improved strategies and policies. Just as in engineering systems, analyzing solutions to a system's governing equations provides insights into the underlying dynamics, enabling design of optimal interventions[51].

At their core, the governing equations approximate the dynamics of interacting shockwaves, representing spikes or waves in community linking activity over time. This minimal framework of coupled linear differential equations provides an analytically tractable model that matches empirical data.

The SI sec. 4 derives these governing dynamical equations (Suppl. Eq. 102) starting from a realistic online grouping mechanism[52,53]. Notably, the equations reproduce the shockwave patterns observed in the link creation data (Figs. 1d and 2). This is because they are mathematically equivalent to shockwave equations—even in their minimal form (Fig. 2):

$$\dot{S}_i = H(t - t_i)\left[a_i(S_{i,0} - S_i) + \sum_j b_{i,j}(S_j - S_i)\right] \tag{1}$$

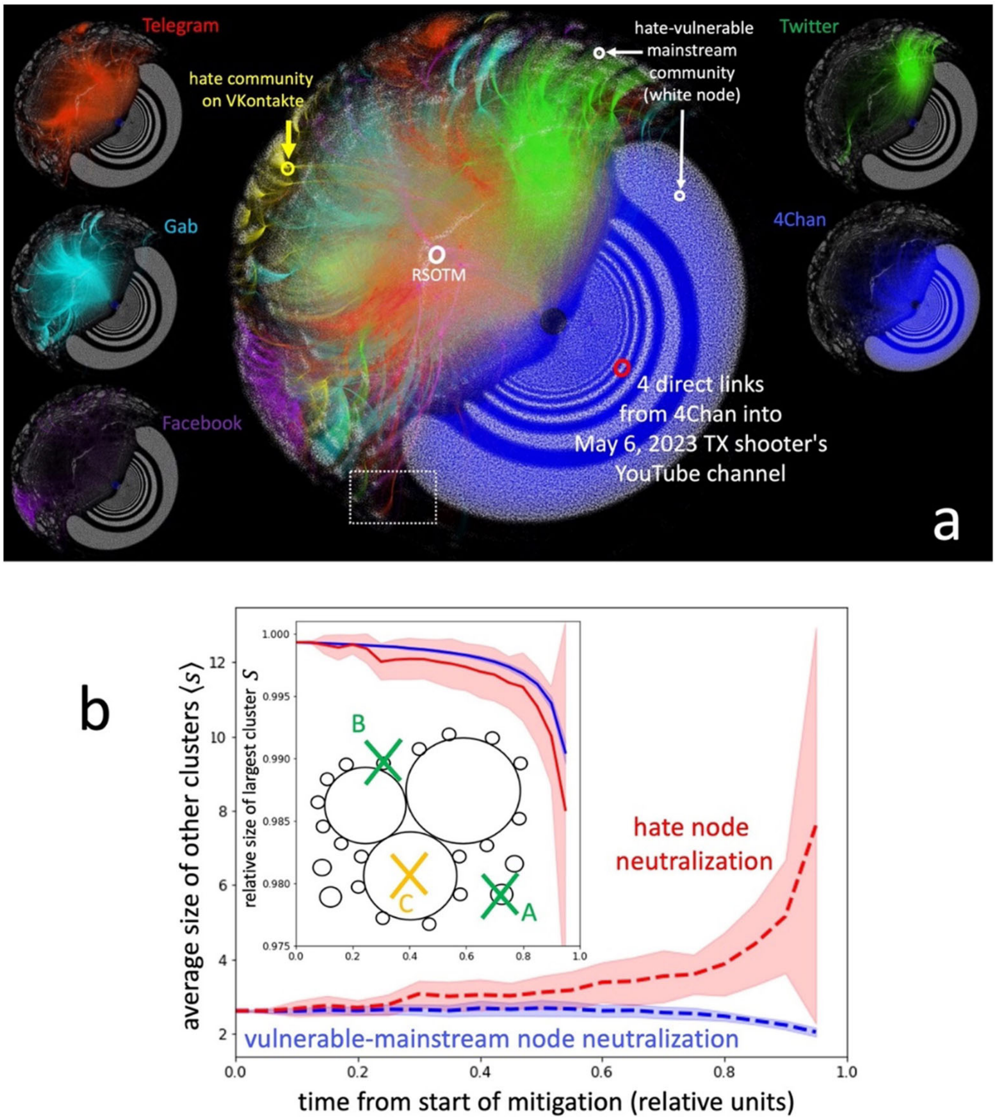

Fig. 3 | Simulations with context of a network map. a Dynamical network aggregated over 2.5 years' worth of social media data. Each colored node is a hate community. Each white node is a hate-vulnerable mainstream community to which a hate node has a direct link. Edges are colored the same color as their source node. Generally, the areas of color visible in this network are space between nodes filled by dense edges. Side panels compare platforms' involvement by only highlighting edges that originate from hate nodes on a given platform. 2023 Texas shooter's YouTube community is shown, so too is a major Wagner mercenary community on Telegram (Reverse Side of the Medal). See SI for others and an example zoom-in (dotted box, Suppl. Fig. 15). Network layout is generated by the ForceAtlas2 algorithm[68]: sets of communities appear closer together when they share more links. **b** Comparing mitigation schemes' effects on network topology, modeled after robustness tests carried out in ref. 55 This simulation neutralizes (removes) a random node at each timestep. The blue line shows this process carried out for hate-vulnerable nodes, while the red line shows this for nodes in the hate core. The main plot shows the simulations' effect on the average size of non-largest connected components, while the inset plot shows the effects on the largest connected component. X-axis is scaled by the total time to neutralize all nodes. Inset uses schematic of the multi-platform network to explain these different curve behaviors: neutralizing nodes in clusters like A (disconnected from the largest connected component) has little effect on S but decreases $\langle s \rangle$; neutralizing nodes in clusters like B can slightly decrease S but increase $\langle s \rangle$; neutralizing nodes in clusters like C can strongly decrease S and strongly increase $\langle s \rangle$.

Here, $H(\ldots)$ represents the Heaviside function, and t_i denotes the onset time of a new wave of link creation (SI sec. 4). The term $a_i\left(S_{i,0} - S_i\right)$ characterizes the intrinsic growth dynamics of community i where a_i $b_{i,j}S_{i,0}$ is the carrying capacity. $\Sigma_j b_{i,j}(S_j - S_i)$ describes the coupling between communities, with $b_{i,j}$ quantifying the functional dependence of the trajectory of community i on that of community j. As analyzed in SI Sec. 4.1.5.1, setting most $b_{i,j}$ terms to zero provides a balance of model simplicity and good fit to the data according to quantitative goodness-of-fit assessments. S_i and S_j capture the time evolution for communities i and j, respectively. Critically, these equations are exactly piecewise solvable in their approximate form (SI Sec. 4.1.5) which means there are no computational errors or instabilities in their solutions and hence predictions.

Figure 2 demonstrates, contrary to common assumptions, Twitter does not act as a central driving platform. Rather, the collective activity of smaller platforms emerges as the key driver, while Twitter serves more as an event reporter. Further, Fig. 2c, d demonstrate the flexibility of those same governing equations by showing that they can reproduce the link-dynamics on a more normal date range, January 4–5, with appropriately lower amplitudes. This suggests the interpretation that the January 6 events represented an amplified version of normal background hate network activity and dynamics, rather than an entirely unpredictable 'Black Swan' event.

The deterministic nature of the governing equations, operating at the many-link-node level, allows for quantitative forecasting at the more general meso- and macro-scales. For example, forward iteration can predict the hourly effects of a proposed mitigation strategy on future ecosystem dynamics. Conversely, the process can be reversed to identify the best mitigation that achieves a desired impact. These findings thus demonstrate how our analysis provides abiding insights and quantitative predictions.

Furthermore, these dynamical link equations provide a tipping point condition for whether hate content can spread system-wide, and how to prevent it. SI Sec. 4.2 derives this: here for simplicity, assume a community digests new hate content received via a link after time $T_{\text{digest}-\text{hate}}$ on average; subsequently forgets it after time $T_{\text{forget}-\text{hate}}$ on average; links between community clusters arise after time $T_{\text{create}-\text{links}}$ on average; and links disappear after time $T_{\text{lose}-\text{links}}$ on average. Then Suppl. Eq. 131 predicts system-wide propagation of hate content will be impeded if

$$T_{\text{lose}-\text{links}}T_{\text{forget}-\text{hate}}(T_{\text{create}-\text{links}}T_{\text{digest}-\text{hate}})^{-1} < 1\ (\text{Suppl.Eq.131})$$

which agrees with simulations. Therefore, system-wide propagation is impeded by prolonging the time to create links or digest hate content, or alternatively, by shortening the time to eliminate links or forget hate content, so that the inequality condition holds.

Mathematically, this confirms why criteria (1)–(3) are crucial, and why mitigation or legislation that does not satisfy them will prove ineffective.

Hate Network Landscape

Aggregating the link dynamics over time (Fig. 3a) shows even more clearly that online hate does not live at the fringe, and that large platforms are not the key. The many smaller platforms in Fig. 3a act like dynamical glue that binds the hate networks together and attaches them directly to the mainstream.

Real-world events involving hate are mirrored—and may increasingly be pre-empted—by hate activity within this dynamical network. In addition to the events in Figs. 1c and 3a shows the 6 May 2023 Texas shooter's YouTube community (now banned) which had attracted 4 separate links into it from other hate communities prior to his attack. Members of these hate communities had been alerted to his YouTube channel and could have easily posted comments and/or content that fueled extreme views and hence influence among his channel's members—including him.

Mentions of RWDS ('Right Wing Death Squad') are also prevalent across Fig. 3a; so too are Wagner mercenary communities (see Suppl. Figs. 7–9). "RWDS" has appeared as insignia worn by recent mass shooters and members of neo-Nazi units in the Ukraine-Russia conflict. Figure 3a also reveals how some hate-vulnerable communities have far higher exposure risk than others – not necessarily because of their views, but because they are more appealing prey. Regardless of the reason for their popularity in hate communities, they attract more links and so sit closer to the hate core in Fig. 3a because of its node-repulsion-link-attraction ForceAtlas2 layout.

This also explains its ordered circles akin to solar system orbital structure: successive subsets of hate-vulnerable nodes have 1, 2, 3, etc. links from 4Chan (blue) and hence a net spring force pulling them toward the hate core that is 1, 2, 3, etc. times as strong. These successively smaller radius stripes hence contain hate-vulnerable communities that are roughly 1, 2, 3, etc. times more likely to receive hate content and influence.

This suggests tailoring pre-emptive action first on the inner rings closest to the hate core in Fig. 3a, then working in order of increasing radius and hence decreasing risk of exposure.

Discussion

We have shown how our approach has led to a deeper understanding of the machinery and mechanisms that manage to sustain online hate at scale. By mapping out the dynamical structure of the online hate network across platforms, we found that its dynamical features contradict current thinking. Instead of the large platforms being the key drivers, hate networks are bound together over time by waves of adaptive links across numerous smaller platforms. This allows the hate networks to progressively strengthen, sidestep mitigations, and exert greater direct influence on the massive mainstream neighbor.

We then derived deterministic governing equations from first principles to describe the link dynamics sustaining online hate. The equations are built on models of human grouping behavior, providing a framework akin to conventional engineering systems. This enables quantitative predictions about recent and future events at hourly resolution, offering enduring insights.

Such knowledge of the dynamical network and its underlying equations allows rigorously calculating and comparing expected impacts of different mitigation strategies. Formal control theory can now be leveraged to systematically design interventions[54].

Figure 3b uses simulations to compare mitigation variants. These simulations randomly remove nodes either found in the set of hate communities (red) or the set of hate-vulnerable communities (blue). In such mitigation strategies, posts are removed from a community if they link to extreme content (e.g., a hate manifesto or footage from a mass shooting) posted in other communities. Iterating this continually with slow link dynamics mimics neutralizing communities across the network shown in Fig. 3a.

The simulated mitigation impacts over time (Fig. 3b) are unlike prior estimates and again challenge current thinking. Figure 3b inset shows the impact on the relative size of the largest cluster (S) which, in a slow link-dynamic limit, represents the maximum spread that any piece of hateful content can have. The main panel shows the impacts on the average size $\langle s\rangle$ of the other clusters which quantifies how linked the remaining communities are on average, and hence quantifies the threat they pose as nucleation sites for further activities.

These curves differ markedly from exponential and scale-free network models associated with the World Wide Web or Internet as a whole[55]. No universally optimal mitigation emerges from our simulations: removing hate nodes irrespective of platform reduces S quicker than removing hate-vulnerable nodes (inset: red curve is lower than blue), but it has the disadvantage that it generates larger nucleation clusters (larger $\langle s\rangle$) for future harms (main: red curve is higher than blue). In terms of the three criteria outlined previously, this means that while such a strategy addresses (1) and (3) by focusing on the system level link-spreading infrastructure rather than individual platforms or links, it may not adequately address (2) – the necessity for fast-paced measures around large "waves" of links – depending on the composition of the smaller nucleation clusters which factor into $\langle s\rangle$. In this area, further study of the interactions between large and small platforms within the context of the hate ecosystem is necessary.

While AI's future weaponization remains uncertain[56], our framework clarifies the dynamics of this online machinery. Future improvements will include: (1) subclassifying hate by type (e.g., anti-Semitic); (2) exploring other hate definitions; (3) analyzing blended hate types; (4) sub-classifying hate-vulnerable communities; (5) incorporating private platforms; (6) adding link weights according their use; (7) adding mainstream-to-hate communities links; (8) adding links to government sources (e.g. Fig. S3); (9) include general harm types; (10) subclassifying each community by location or scale (e.g. local vs. global[57,58]).

Although our data is technically a large sample of the unknown true online population, the billions of estimated users suggest it qualifies as a crude population map. Moreover, we charted this hate ecosystem not by isolated sampling but by algorithmically tracing links node-to-node. This process tended to eventually return to the same nodes and hence, like circumnavigating the globe, it hints that we have charted out—albeit crudely—the skeleton of the true online hate ecosystem.

Methods

Data collection and link tracking

Our methodology follows refs. 59,60 (SI Sec. 1) but goes beyond prior studies by (i) including the mainstream communities that hate communities link to over time (referred to in this paper as "hate-vulnerable"), (ii) tracking this data down to second-scale resolution across 13 platforms, (iii) including new decentralized[61] and blockchain platforms (e.g. Minds, Steemit) for which blame cannot be pinned on single servers and cryptocurrency can incentivize users, and (iv) including gaming-related platforms[62] such as Discord which played a key role in recent security leaks.

Our focus is on platform-provided, built-in communities because people join these to develop their shared interests[63–66] including hate. Terminology for such communities is platform-specific. Examples are a VKontakte Club (VKontakte is a social media platform controlled by Russian state-owned bank Gazprombank and insurance company Sogaz[67]); a Facebook Page; a Telegram Channel; a Gab Group. Each community contains anywhere from a few to a few million users and is unrelated to "community" as used in network community detection.

A "hate" community is one in which 2 or more of its 20 most recent posts include U.S. Department of Justice-defined hate speech. Together, these hate communities make up the "hate core". A "hate-vulnerable" community is one that is outside this hate community core but was linked to directly by a hate community (Fig. 1a inset). Hate-vulnerable communities' views can vary significantly (we do not attempt to categorize them for this study), but mostly represent a benign mainstream that have become targets of the hate core (SI Sec. 1.2). Concretely, any link in this network appears because a post included a URL directly to another community or to content featured therein. For example, one edge in the network in Fig. 3a represents the URL to a Gab post posted in a comment thread on 4chan. See this and other examples in Fig. S5.

A link to community B can appear in community A at some time t if B's content is of interest to A's members (Suppl. Figs. 1 and 4 show examples). The link directs attention of A's members to B, which may be on a different platform and another language. A's members can then add comments and content on B without B's members knowing this incoming link exists. Hence B's members can unwittingly experience direct exposure to, and influence from, A's hateful narratives. Since our focus is on hate networks, we do not include links originating in hate-vulnerable nodes.

No individual information is required. Only public communities are accessed, but the ecosystem of open communities provides a skeleton on which private communities sit (Suppl. Fig. 3).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets used in this study contain sensitive information from social media platforms. To comply with data protection standards and avoid potential misuse, the raw data cannot be shared publicly; however, the preprocessed derivative datasets which can be used to reproduce the results in the study are available in our Data Access repository, https://github.com/gwdonlab/data-access.

Code availability

Figure 1 was created using R (plotly and sankey3D packages), the latter package can be accessed at: https://github.com/fbreitwieser/sankeyD3; Fig. 3a was created using Gephi, an open-source software available at https://gephi.org. The simulations in Fig. 2 were performed using proprietary software from Wolfram Research, the implementation and code of which is described at https://community.wolfram.com/groups/-/m/t/2981192. The additional network detailed data is given at https://github.com/gwdonlab/data-access. Together, this provides readers with access to the minimum dataset that is necessary to interpret, verify, and extend the research in the article.

Received: 28 August 2023; Accepted: 24 November 2023;
Published online: 17 April 2024

References

1. Brown, R. & Livingston, L. A New Approach to Assessing the Role of Technology in Spurring and Mitigating Conflict: Evidence From Research and Practice. *JIA SIPA* https://jia.sipa.columbia.edu/new-approach-assessing-role-technology-spurring-and-mitigating-conflict-evidence-research-and (2018).
2. Starbird, K. Disinformation's spread: bots, trolls and all of us. *Nature* **571**, 449–449 (2019).
3. Lamensch, M. To eliminate violence against women, we must take the fight to online spaces. *Centre for International Governance Innovation* https://www.cigionline.org/articles/to-eliminate-violence-against-women-we-must-take-the-fight-to-online-spaces/ (2022)
4. Illegal trade in AI child sex abuse images exposed. *BBC News* (2023).
5. Surge in young children being targeted by cyber bullies. *The West Australian* https://thewest.com.au/news/social/surge-in-young-children-being-targeted-by-cyber-bullies-c-11223220 (2023).
6. Gill, P. & Corner, E. Lone-Actor Terrorist Use of the Internet & Behavioural Correlates. In *Terrorism Online: Politics, Law, Technology and Unconventional Violence* (eds. L. Jarvis, S. Macdonald, & T. Chen) (Routledge, 2015).
7. Cynthia Miller-Idriss. *Hate in the Homeland: The New Global Far Right*. (2020).
8. The haters and conspiracy theorists back on Twitter. *BBC News* (2023).
9. DiResta, R. The Digital Maginot Line. *ribbonfarm* https://www.ribbonfarm.com/2018/11/28/the-digital-maginot-line/ (2018).
10. Vesna, C.-G. & Maslo-Čerkić, Š. *Hate speech online and the approach of the Council of Europe and the European Union*. https://urn.nsk.hr/urn:nbn:hr:118:377009 (2023).
11. House of Commons Home Affairs Committee. *14th Report - Hate crime: abuse, hate and extremism online*. 1–34 https://publications.parliament.uk/pa/cm201617/cmselect/cmhaff/609/609.pdf (2017).
12. Hart, R. White Supremacist Propaganda Hit Record Levels In 2022, ADL Says. *Forbes* https://www.forbes.com/sites/roberthart/2023/03/09/white-supremacist-propaganda-hit-record-levels-in-2022-adl-says/ (2023).
13. Online Hate and Harassment: The American Experience. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2023 (2023).
14. Yael Eisenstat. Hate is surging online — and social media companies are in denial. Congress can help protect users. *The Hill* https://thehill.com/opinion/congress-blog/4085909-hate-is-surging-online-and-social-media-companies-are-in-denial-congress-can-help-protect-users/ (2023).
15. Nelson, D. / J. UN Warns of AI-Generated Deepfakes Fueling Hate and Misinformation Online. *Decrypt* https://decrypt.co/144281/un-united-nations-ai-deepfakes-hate-misinformation (2023).
16. United Nations. *Common Agenda Policy Brief: Information Integrity on Digital Platforms*. https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf (2023).
17. Douek, E. Content moderation as systems thinking. *Harvard Law Review.* Vol. 136 (2022).
18. Coping with the fear of mass shootings. *SiouxlandProud | Sioux City, IA | News, Weather, and Sports* https://www.siouxlandproud.com/news/local-news/coping-with-the-fear-of-mass-shootings/ (2023).
19. One-Third of U.S. Adults Say Fear of Mass Shootings Prevents Them From Going to Certain Places or Events. https://www.socialworktoday.com/news/dn_081519.shtml (2021).
20. Milmo, D. & Hern, A. Elections in UK and US at risk from AI-driven disinformation, say experts. *The Guardian* (2023).
21. Hsu, T. & Myers, S. L. A.I.'s Use in Elections Sets Off a Scramble for Guardrails. *The New York Times* (2023).
22. Aut, N., Ranaware, S., Ghadge, S., Jadhav, R. & Jagtap, P. Social media based hate speech detection using machine learning. *IJRASET*. Vol. 11 (2023).
23. Ollagnier, A., Cabrio, E. & Villata, S. Harnessing Bullying Traces to Enhance Bullying Participant Role Identification in Multi-Party Chats. *The International FLAIRS Conference Proceedings.* Vol. 36 (2023).

24. Aldreabi, E., Lee, J. M. & Blackburn, J. Using Deep Learning to Detect Islamophobia on Reddit. *The International FLAIRS Conference Proceedings.* Vol. 36 (2023).
25. Morgan, M. & Kulkarni, A. Platform-agnostic Model to Detect Sinophobia on Social Media. In *Proceedings of the 2023 ACM Southeast Conference* 149–153 (Association for Computing Machinery, 2023).
26. Beacken, G., Trauthig, I. & Woolley, S. Platforms' Efforts to Block Antisemitic Content Are Falling Short. *Centre for International Governance Innovation* https://www.cigionline.org/articles/platforms-efforts-to-block-anti-semitic-content-are-falling-short/ (2022).
27. Cinelli, M. et al. Dynamics of online hate and misinformation. *Sci. Rep.* **11**, 22083 (2021).
28. Chen, E., Lerman, K. & Ferrara, E. Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus twitter data set. *JMIR Public Health Surveillance* **6**, e19273 (2020).
29. Gelfand, M. J., Harrington, J. R. & Jackson, J. C. The strength of social norms across human groups. *Perspect. Psychol. Sci.* **12**, 800–809 (2017).
30. van der Linden, S., Leiserowitz, A., Rosenthal, S. & Maibach, E. Inoculating the public against misinformation about climate change. *Glob. Challenges* **1**, 1600008 (2017).
31. Lewandowsky, S. et al. *Debunking Handbook 2020*. 1–19 https://www.climatechangecommunication.org/wp-content/uploads/2020/10/DebunkingHandbook2020.pdf (2020).
32. Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094–1096 (2018).
33. Smith, R., Cubbon, S. & Wardle, C. *Under the surface: Covid-19 vaccine narratives, misinformation and data deficits on social media*. https://firstdraftnews.org/long-form-article/under-the-surface-covid-19-vaccine-narratives-misinformation-and-data-deficits-on-social-media/ (2020).
34. Semenov, A. et al. Exploring Social Media Network Landscape of Post-Soviet Space. *IEEE Access* **7**, 411–426 (2019).
35. Rao, A., Morstatter, F. & Lerman, K. Partisan asymmetries in exposure to misinformation. *Sci. Rep.* **12**, 15671 (2022).
36. Wu, X.-Z., Fennell, P. G., Percus, A. G. & Lerman, K. Degree correlations amplify the growth of cascades in networks. *Phys. Rev. E* **98**, 022321 (2018).
37. Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S. & Lewandowsky, S. Psychological inoculation improves resilience against misinformation on social media. *Sci. Adv.* **8**, eabo6254 (2022).
38. Green, Y. et al. *Evidence-Based Misinformation Interventions: Challenges and Opportunities for Measurement and Collaboration*. https://carnegieendowment.org/2023/01/09/evidence-based-misinformation-interventions-challenges-and-opportunities-for-measurement-and-collaboration-pub-88661 (2023).
39. Dynamic Online Networks Lab. *Literature Review*. https://bpb-us-e1.wpmucdn.com/blogs.gwu.edu/dist/5/3446/files/2022/10/lit_review.pdf.
40. DisinfoDocket. *DisinfoDocket* https://www.disinfodocket.com/ (2024).
41. DisinfoDocket 12 July. *DisinfoDocket* https://www.disinfodocket.com/dd-12jul23/ (2023).
42. Strengthened Code of Practice on Disinformation: Signatories to identify ways to step up work one year after launch | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/strengthened-code-practice-disinformation-signatories-identify-ways-step-work-one-year-after-launch (2023).
43. The Artificial Intelligence Act. *The Artificial Intelligence Act* https://web.archive.org/web/20230811085634/. https://artificialintelligenceact.eu/ (2021).
44. Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines | Shaping Europe's digital future. https://digital-strategy.ec.europa.eu/en/news/digital-services-act-commission-designates-first-set-very-large-online-platforms-and-search-engines (2023).
45. Benninger, M. 'Fringe' websites radicalized Buffalo shooter, report concludes. https://www.wbng.com. https://www.wbng.com/2022/10/18/fringe-websites-radicalized-buffalo-shooter-report-concludes/ (2022).
46. Fringe Social Media: Are you digging deep enough? | SMI Aware. https://web.archive.org/web/20201001222412/. https://smiaware.com/blog/fringe-social-media-are-you-digging-deep-enough/ (2019).
47. Rodrigo, C. M. & Klar, R. Fringe social networks boosted after mob attack. *The Hill* https://thehill.com/policy/technology/533919-fringe-social-networks-boosted-after-mob-attack/ (2021).
48. Hsu, T. News on Fringe Social Sites Draws Limited but Loyal Fans, Report Finds. *The New York Times* (2022).
49. Reporter, C. D. N. S. On fringe social media sites, Buffalo mass shooting becomes rallying call for white supremacists. *Buffalo News.* https://buffalonews.com/news/local/on-fringe-social-media-sites-buffalo-mass-shooting-becomes-rallying-call-for-white-supremacists/article_74a55388-f61b-11ec-812a-97d8f2646d45.html (2022).
50. Fringe social media networks sidestep online content rules. *POLITICO* https://www.politico.eu/article/fringe-social-media-telegram-extremism-far-right/ (2022).
51. Gavrilets, S. Collective action and the collaborative brain. *J. R. Soc. Interface* **12**, 20141067 (2015).
52. Forsyth, D. R. *Group Dynamics*. (Wadsworth Cengage Learning, 2014).
53. Palla, G., Barabási, A.-L. & Vicsek, T. Quantifying social group evolution. *Nature* **446**, 664–667 (2007).
54. Liu, Y.-Y. & Barabási, A.-L. Control principles of complex systems. *Rev. Mod. Phys.* **88**, 035006 (2016).
55. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
56. Sear, R. F., Leahy, R., Restrepo, N. J., Lupu, Y. & Johnson, N. F. Dynamic Latent Dirichlet Allocation Tracks Evolution of Online Hate Topics. *Adv. Artif. Intell. Mach. Learn.* **2**, 257–272 (2022).
57. Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A. & Danforth, C. M. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. *PLoS ONE* **6**, e26752 (2011).
58. Ball, P. Hot Air by Peter Stott review – the battle against climate change denial. *The Guardian* (2021).
59. Lupu, Y. et al. Offline events and online hate. *PLoS ONE* **18**, e0278511 (2023).
60. Velásquez, N. et al. Online hate network spreads malicious COVID-19 content outside the control of individual social media platforms. *Sci Rep* **11**, 11549 (2021).
61. Nix, N. Meta considers a new social network, as decentralized model gains steam. *Washington Post* (2023).
62. *Hate Is No Game: Hate and Harassment in Online Games*. https://www.adl.org/resources/report/hate-no-game-hate-and-harassment-online-games-2022 (2022).
63. Ammari, T. & Schoenebeck, S. "Thanks for your interest in our Facebook group, but it's only for dads": Social Roles of Stay-at-Home Dads. in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* 1363–1375 (Association for Computing Machinery, 2016).
64. Moon, R. Y., Mathews, A., Oden, R. & Carlin, R. Mothers' perceptions of the internet and social media as sources of parenting and health information: qualitative study. *J. Med. Internet Res.* **21**, e14289 (2019).

65. Laws, R. et al. Differences between mothers and fathers of young children in their use of the internet to support healthy family lifestyle behaviors: cross-sectional study. *J. Med. Int. Res.* **21**, e11454 (2019).
66. Madhusoodanan, J. Safe space: online groups lift up women in tech. *Nature* **611**, 839–841 (2022).
67. Times, T. M. Gazprom Gains Control of Russia's Top Social Network. *The Moscow Times* https://www.themoscowtimes.com/2021/12/03/gazprom-gains-control-of-russias-top-social-network-a75724 (2021).
68. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE* **9**, e98679 (2014).

Acknowledgements

N.F.J. is supported by U.S. Air Force Office of Scientific Research awards FA9550-20-1-0382 and FA9550-20-1-0383, as well as the John Templeton Foundation.

Author contributions

R.S. and N.J.R. contributed to obtaining, organizing, and preserving the data resources. M.Z. analyzed the results and generated Fig. 3. L.I. generated Figs. 1 and 2. N.F.J. supervised the project. N.F.J. wrote the paper. All authors were involved in reviewing the final manuscript, and in the conceptualization, methodology, and validation. All correspondence and material requests should be addressed to N.F.J. neiljohnson@gwu.edu.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44260-024-00002-2.

Correspondence and requests for materials should be addressed to Neil F. Johnson.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.